Amendment to the Claims

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1. (currently amended) A system for grouping clusters of semantically scored documents electronically stored in a data corpus, comprising:

a scoring module determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight based on a number of terms for the at least one concept, a structural weight, and a corpus weight, forming the score assigned to the at least one concept as a normalized score vector for each such document, and determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

a clustering module forming clusters of the documents, comprising:

a selection submodule selecting a set of candidate seed documents from the plurality of documents;

a seed document identification submodule identifying a set of seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

a non-seed document identification submodule identifying a plurality of non-seed documents;

a comparison submodule determining the similarity between each non-seed document and a center of each cluster; and

a clustering submodule grouping each such non-seed document into a cluster with a best fit, subject to a minimum fit;

a threshold module determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document, dynamically determining a threshold for each cluster as a function of the similarity between each of the documents, and identifying and reassigning each of the documents having the similarity falling outside the threshold.

Claim 2 (canceled).

3. (previously presented) A system according to Claim 1, further comprising:

a compression module compressing the score through logarithmic compression.

Claim 4 (canceled).

5. (original) A system according to Claim 1, further comprising:

the scoring module calculating the structural weight as a function of a location of the at least one concept within the at least one such document.

6. (original) A system according to Claim 1, further comprising:

the scoring module calculating the corpus weight as a function of a reference count of the at least one concept over the plurality of documents.

Claims 7-8 (canceled).

9. (currently amended) A method for grouping clusters of semantically scored documents electronically stored in a data corpus, comprising:

determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight based on a number of terms for the at least one concept, a structural weight, and a corpus weight;

forming the score assigned to the at least one concept as a normalized score vector for each such document;

determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

forming logically-grouped clusters of the documents, comprising:

selecting a set of candidate seed documents from the plurality of documents;

identifying a set of seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

identifying a plurality of non-seed documents;

determining the similarity between each non-seed document and a center of each cluster; and

grouping each such non-seed document into a cluster with a best fit, subject to a minimum fit;

determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document;

dynamically determining a threshold for each cluster as a function of the similarity between each of the documents; and

identifying and reassigning each of the documents having the similarity falling outside the threshold.

Claim 10 (canceled).

11. (previously presented) A method according to Claim 9, further comprising:

compressing the score through logarithmic compression.

Claim 12 (canceled).

13. (original) A method according to Claim 9, further comprising:

calculating the structural weight as a function of a location of the at least one concept within the at least one such document.

14. (original) A method according to Claim 9, further comprising:

calculating the corpus weight as a function of a reference count of the at least one concept over the plurality of documents.

Claims 15-16 (canceled).

17. (currently amended) A computer-readable storage medium holding code for grouping clusters of semantically scored documents electronically stored in a data corpus, comprising:

code for determining a score, which is assigned to at least one concept that has been extracted from a plurality of electronically-stored documents, wherein the score is calculated as a function of a summation of a frequency of occurrence of the at least one concept within at least one such document, a concept weight based on a number of terms for the at least one concept, a structural weight, and a corpus weight;

code for forming the score assigned to the at least one concept as a normalized score vector for each such document;

code for determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

code for forming logically-grouped clusters of the documents, comprising;

code for selecting a set of candidate seed documents from the plurality of documents;

code for identifying a set of seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

code for identifying a plurality of non-seed documents;

code for determining the similarity between each non-seed document and a center of each cluster; and

code for grouping each such non-seed document into a cluster with a best fit, subject to a minimum fit;

code for determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document;

code for dynamically determining a threshold for each cluster as a function of the similarity between each of the documents; and

code for identifying and reassigning each of the documents having the similarity falling outside the threshold.

18. (currently amended) A system for providing efficient document scoring of concepts within and clustering of documents in an electronically-stored document set, comprising:

a scoring module scoring a document in an electronically-stored document set, comprising:

a frequency module determining a frequency of occurrence of at least one concept within a document;

a concept weight module analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document, wherein the concept weight is based on a number of terms for the at least one concept;

a structural weight module analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

a corpus weight module analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document;

a scoring evaluation module evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight;

a vector module forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set; and

a determination module determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

a clustering module grouping the documents by the score into a plurality of clusters, comprising:

a selection submodule selecting a set of candidate seed documents from the electronically-stored document set;

a cluster seed submodule identifying seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

an identification submodule identifying a plurality of non-seed documents;

a comparison submodule determining the similarity between each non-seed document and a cluster center of each cluster; and

a clustering submodule assigning each non-seed document to the cluster with a best fit, subject to a minimum fit; and

a threshold module relocating outlier documents, comprising determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document, dynamically determining a threshold for each cluster as a function of the similarity between each of the documents, and identifying and reassigning each of the documents with the similarity falling outside the threshold.

19. (previously presented) A system according to Claim 18, further comprising:

the scoring module evaluating the score in accordance with the formula:

$$S_i = \sum_{1 \to n}^{j} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$ comprises the corpus weight for occurrence j of concept i.

20. (currently amended) A system according to Claim 19, further comprising:

the concept weight module evaluating the concept weight in accordance with the formula:

$$cw_{ij} = \begin{cases} 0.25 + (0.25 \times t_{ij}), & 1 \leq t_{ij} \leq 3 \\ 0.25 + (0.25 \times [7 - t_{ij}]), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

where cw_{ij} comprises the concept weight and t_{ij} comprises [[a]] the number of terms for occurrence j of each such concept i.

21. (previously presented) A system according to Claim 19, further comprising:

the structural weight module evaluating the structural weight in accordance with the formula:

$$sw_{ij} = \begin{cases} 1.0, & if(j \approx SUBJECT) \\ 0.8, & if(j \approx HEADING) \\ 0.7, & if(j \approx SUMMARY) \\ 0.5 & if(j \approx BODY) \\ 0.1 & if(j \approx SIGNATURE) \end{cases}$$

where sw_{ij} comprises the structural weight for occurrence j of each such concept i.

22. (previously presented) A system according to Claim 19, further comprising:

the corpus weight module evaluating the corpus weight in accordance with the formula:

$$rw_{ij} = \begin{cases} \left(\dfrac{T - r_{ij}}{T}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for occurrence j of each such concept i, T comprises a total number of reference counts of documents in the document set, and M comprises a maximum reference count of documents in the document set.

23. (previously presented) A system according to Claim 19, further comprising:

a compression module compressing the score in accordance with the formula:

$$S_i' = \log(S_i + 1)$$

where S_i' comprises the compressed score for each such concept i.

24. (original) A system according to Claim 18, further comprising:

a global stop concept vector cache maintaining concepts and terms; and

a filtering module filtering selection of the at least one concept based on the concepts and terms maintained in the global stop concept vector cache.

25. (original) A system according to Claim 18, further comprising:

a parsing module identifying terms within at least one document in the document set, and combining the identified terms into one or more of the concepts.

26. (original) A system according to Claim 25, further comprising:

the parsing module structuring each such identified term in the one or more concepts into canonical concepts comprising at least one of word root, character case, and word ordering.

27. (original) A system according to Claim 25, wherein at least one of nouns, proper nouns and adjectives are included as terms.

Claims 28-30 (canceled).

31. (previously presented) A system according to Claim 18, further comprising:

the similarity submodule calculating the similarity in accordance with the formula:

$$\cos\sigma_{AB} = \frac{\left\langle \vec{S}_A \cdot \vec{S}_B \right\rangle}{\left|\vec{S}_A\right|\left|\vec{S}_B\right|}$$

where $\cos\sigma_{AB}$ comprises a similarity between a document A and a document B, $\vec{S}_A$ comprises a score vector for document A, and $\vec{S}_B$ comprises a score vector for document B.

Claims 32-34 (canceled).

35. (currently amended) A method for providing efficient document scoring of concepts within and clustering of documents in an electronically-stored document set, comprising:

scoring a document in an electronically-stored document set, comprising:

determining a frequency of occurrence of at least one concept within a document;

analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document, wherein the concept weight is based on a number of terms for the at least one concept;

analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight;

forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set;

determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

grouping the documents by the score into a plurality of clusters, comprising:

selecting a set of candidate seed documents from the electronically-stored document set;

identifying seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

identifying a plurality of non-seed documents;

determining the similarity between each non-seed document and a center of each cluster; and

assigning each non-seed document to the cluster with a best fit, subject to a minimum fit; and

relocating outlier documents, comprising:

determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document;

dynamically determining a threshold for each cluster as a function of the similarity between each of the documents; and identifying and reassigning each of the documents with the similarity falling outside the threshold.

36. (previously presented) A method according to Claim 35, further comprising:

evaluating the score in accordance with the formula:

$$S_i = \sum_{1 \to n}^{j} f_{ij} \times cw_{ij} \times sw_{ij} \times rw_{ij}$$

where S_i comprises the score, f_{ij} comprises the frequency, $0 < cw_{ij} \leq 1$ comprises the concept weight, $0 < sw_{ij} \leq 1$ comprises the structural weight, and $0 < rw_{ij} \leq 1$ comprises the corpus weight for occurrence j of concept i.

37. (currently amended) A method according to Claim 36, further comprising:

evaluating the concept weight in accordance with the formula:

$$cw_{ij} = \begin{cases} 0.25 + \left(0.25 \times t_{ij}\right), & 1 \leq t_{ij} \leq 3 \\ 0.25 + \left(0.25 \times \left[7 - t_{ij}\right]\right), & 4 \leq t_{ij} \leq 6 \\ 0.25, & t_{ij} \geq 7 \end{cases}$$

where cw_{ij} comprises the concept weight and t_{ij} comprises [[a]] the number of terms for occurrence j of each such concept i.

38. (previously presented) A method according to Claim 36, further comprising:

evaluating the structural weight in accordance with the formula:

$$sw_{ij} = \begin{cases} 1.0, & if\left(j \approx SUBJECT\right) \\ 0.8, & if\left(j \approx HEADING\right) \\ 0.7, & if\left(j \approx SUMMARY\right) \\ 0.5 & if\left(j \approx BODY\right) \\ 0.1 & if\left(j \approx SIGNATURE\right) \end{cases}$$

where sw_{ij} comprises the structural weight for occurrence j of each such concept i.

39. (previously presented) A method according to Claim 36, further comprising:

evaluating the corpus weight in accordance with the formula:

$$rw_{ij} = \begin{cases} \left(\dfrac{T - r_{ij}}{T}\right)^2, & r_{ij} > M \\ 1.0, & r_{ij} \leq M \end{cases}$$

where rw_{ij} comprises the corpus weight, r_{ij} comprises a reference count for occurrence j of each such concept i, T comprises a total number of reference counts of documents in the document set, and M comprises a maximum reference count of documents in the document set.

40. (previously presented) A method according to Claim 36, further comprising:

compressing the score in accordance with the formula:

$$S_i' = \log(S_i + 1)$$

where S_i' comprises the compressed score for each such concept i.

41. (original) A method according to Claim 35, further comprising:

maintaining concepts and terms in a global stop concept vector cache; and

filtering selection of the at least one concept based on the concepts and terms maintained in the global stop concept vector cache.

42. (original) A method according to Claim 35, further comprising:

identifying terms within at least one document in the document set; and

combining the identified terms into one or more of the concepts.

43. (original) A method according to Claim 42, further comprising:

structuring each such identified term in the one or more concepts into canonical concepts comprising at least one of word root, character case, and word ordering.

44. (original) A method according to Claim 42, further comprising: including as terms at least one of nouns, proper nouns and adjectives.

Claims 45-47 (canceled).

48. (previously presented) A method according to Claim 35, further comprising:

calculating the similarity in accordance with the formula:

$$\cos\sigma_{AB} = \frac{\left\langle \vec{S}_A \cdot \vec{S}_B \right\rangle}{\left|\vec{S}_A\right|\left|\vec{S}_B\right|}$$

where $\cos\sigma_{AB}$ comprises a similarity between a document A and a document B, $\vec{S}_A$ comprises a score vector for document A, and $\vec{S}_B$ comprises a score vector for document B.

Claims 49-51 (canceled).

52. (currently amended) A computer-readable storage medium holding code for providing efficient document scoring of concepts within and clustering of documents in an electronically-stored document set, comprising:

code for scoring a document in an electronically-stored document set, comprising:

code for determining a frequency of occurrence of at least one concept within a document;

code for analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document, wherein the concept weight is based on a number of terms for the at least one concept;

code for analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

code for analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

code for evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight;

code for forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set;

code for determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

code for grouping the documents by the score into a plurality of clusters, comprising:

code for selecting a set of candidate seed documents from the electronically-stored document set;

code for identifying seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

code for identifying a plurality of non-seed documents;

code for determining the similarity between each non-seed document and a center of each cluster; and

code for assigning each non-seed document to the cluster with a best fit, subject to a minimum fit; and

code for relocating outlier documents, comprising:

code for determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document;

code for dynamically determining a threshold for each cluster as a function of the similarity between each of the documents; and code for identifying and reassigning each of the documents with the similarity falling outside the threshold.

53. (currently amended) An apparatus for providing efficient document scoring of concepts within and clustering of documents in an electronically-stored document set, comprising:

means for scoring a document in an electronically-stored document set, comprising:

means for determining a frequency of occurrence of at least one concept within a document;

means for analyzing a concept weight reflecting a specificity of meaning for the at least one concept within the document, wherein the concept weight is based on a number of terms for the at least one concept;

means for analyzing a structural weight reflecting a degree of significance based on structural location within the document for the at least one concept;

means for analyzing a corpus weight inversely weighing a reference count of occurrences for the at least one concept within the document; and

means for evaluating a score to be associated with the at least one concept as a function of a summation of the frequency, concept weight, structural weight, and corpus weight;

means for forming the score assigned to the at least one concept as a normalized score vector for each such document in the electronically-stored document set;

means for determining a similarity between the normalized score vector for each such document as an inner product of each normalized score vector;

means for grouping the documents by the score into a plurality of clusters, comprising:

means for selecting a set of candidate seed documents from the electronically-stored document set;

means for identifying seed documents by applying the similarity to each such candidate seed document and selecting those candidate seed documents that are sufficiently unique from other candidate seed documents as the seed documents;

means for identifying a plurality of non-seed documents;

means for determining the similarity between each non-seed document and a center of each cluster; and

means for assigning each non-seed document to the cluster with a best fit, subject to a minimum fit; and

means for relocating outlier documents, comprising:

means for determining the similarity between each of the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the at least one concepts in that document;

means for dynamically determining a threshold for each cluster as a function of the similarity between each of the documents; and

means for identifying and reassigning each of the documents with the similarity falling outside the threshold.